Journal of
Epidemiology and Community Health, 1979, 33, 104-106

Some insights into Miettinen's multivariate confounder score approach to case-control study analysis

M. C. PIKE AND J. ANDERSON
From the Department of Community and Family Medicine, University of Southern California School of Medicine

N. DAY
From the National Cancer Institute, Bethesda

SUMMARY We have studied Miettinen's multivariate confounder score method of controlling confounding in case-control studies both theoretically and by simulation. The main conclusion to be drawn from our results is that the method will in many practical situations seriously exaggerate the statistical significance achieved, and its use is not to be recommended.

In epidemiologic case-control studies the standard approach to control confounding is to cross-classify the subjects by the confounding factors: the method of combining 2 × 2 tables of Mantel and Haenszel (1959) is of this type. The number of strata involved may be large even if there are only a few confounding variables; thus the resulting analysis is often inefficient mainly because of certain strata including no cases or no controls.

A number of multivariate approaches to this problem have been suggested, based on classical statistical ideas. Logistic discriminant analysis is particularly useful (Breslow and Powers, 1978) and it has recently been extended to case-control studies involving matching (Holford *et al.*, 1978).

A non-standard alternative multivariate approach has been suggested by Miettinen (1976) and widely adopted. This approach involves the construction of a 'multivariate confounder score' with a rationale that does not follow from classical statistical principles. The purpose of this paper is to present this method in standard statistical terms.

Confounder score

It is sufficient to consider only case-control studies in which the number of cases is equal to the number of controls. Let this number be n. For each person we have measurements on k variables ($X_1 \ldots X_k$), where X_k is the variable of main interest and ($X_1 \ldots X_{k-1}$) are the possibly confounding variables. Write the observed values of the variables for the *ith* person as ($x_{i1} \ldots x_{ik}$) and let the cases take the n values $i = 1 \ldots n$ and the controls the n values $i = n + 1 \ldots 2n$.

To calculate the multivariate confounder scores for the 2n cases and controls, begin by calculating a standard multivariate normal discriminant analysis between cases and controls (Anderson, 1957). Write the resulting discriminant function as:

$$D = \hat{b}_o + \hat{b}_1 X_1 + \ldots + \hat{b}_k X_k \quad (1)$$

The multivariate confounder score is then taken to be

$$S = \hat{b}_o + \hat{b}_1 X_1 + \ldots + \hat{b}_{k-1} X_{k-1} \quad (2)$$

So the score for person i is

$$S_i = \hat{b}_o + \hat{b}_1 x_{i1} + \ldots + \hat{b}_{k-1} x_{i,\,k-1} \quad (3)$$

The range of values of S_i is then examined and, after excluding cases and controls whose S values lie outside the range of the scores that is common to both cases and controls, a number of nearly equal strata (usually five) are defined, based on the remaining S values. These 'stratified' sets of cases and controls are then analysed for the effect of X_k by standard methods.

The standard regression approach

Standard multivariate discriminant analysis may be expressed in linear regression terms (Anderson, 1957). A regression analysis of Y ($Y = 0{\cdot}5$ if a case and $Y = -0{\cdot}5$ if a control) against ($X_1 \ldots X_k$) will produce a regression equation

$$Y = \hat{c}_o + \hat{c}_1 X_1 + \ldots + \hat{c}_k X_k \quad (4)$$

where the $\hat{c}_j$'s are proportional to the $\hat{b}_j$'s of equation (2).

If we write the calculated regression line of Y against $(X_1 \ldots X_{k-1})$ as

$$Y = \hat{d}_o + \hat{d}_1 X_1 + \ldots + \hat{d}_{k-1} X_{k-1} \qquad (5)$$

then the standard analysis of X_k allowing for $(X_1 \ldots X_{k-1})$ will compare the residual sum of squares associated with (4) with the residual sum of squares associated with (5).

Confounder scores as covariates

Except that certain 'outlier' observations are discarded and the S scores collapsed into strata rather than their distinct values considered, the confounder score method is equivalent to a covariance analysis of Y on X_k with covariate S. The residual sum of squares after fitting the covariance equation

$$Y = A + BS + CX_k \qquad (6)$$

is the same as the residual sum of squares associated with (4), since the least squares fit of (6) will obviously be with $\hat{A} = 0$, $\hat{B} = 1$ and $\hat{C} = \hat{c}_k$ since this is equation (4). This residual sum of squares will be tested for statistical significance by comparison with the residual sum of squares after fitting the equation

$$Y = D + ES \qquad (7)$$

The residual sum of squares associated with (7) will of necessity be greater than the residual sum of squares associated with (5), so that the test of significance for $\hat{C}$ in equation (6) will be incorrect and greater than is warranted, although the actual value of $\hat{C}$ will be correct.

An example and discussion

To understand in what circumstances the errors in the level of significance will be important, consider the situation with $k = 3$ and where the distribution of (X_1, X_2, X_3) is trivariate normal with a simple structure.

Let the mean vector for the cases be (μ_1, μ_2, μ_3) and for the controls (0, 0, 0); and let both cases and controls have the same covariance matrix

$$V = \begin{matrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{matrix}$$

The expectation of X_3 given (X_1, X_2) is

$$E(X_3 \mid X_1, X_2, \text{case}) = \mu_3 + \rho(X_1 - \mu_1) + \rho(X_2 - \mu_2) \qquad (8)$$

and

$$E(X_3 \mid X_1, X_2, \text{control}) = \rho X_1 + \rho X_2 \qquad (9)$$

Thus for the discriminating value of X_3, as evidenced by the difference between μ_3 and 0, to be totally 'explained' by (X_1, X_2), we have to have (8) $\equiv$ (9). Thus if

$$\mu_3 - \rho\mu_1 - \rho\mu_2 = 0$$

$$\text{i.e. } \mu_3 = \rho\mu_1 + \rho\mu_2 \qquad (10)$$

then X_3 has no ability given (X_1, X_2) to discriminate further between cases and controls.

We have investigated by simulation the consequences of using the multivariate confounder score approach (as given in the section above) to data generated using this model with equation (10) holding. Each situation simulated had $\mu_1 = \mu_2$ and the results are shown in the Table. The first column is the value of ρ for the generating covariance matrix. The remaining columns show for four nominal 1-sided significance levels the estimate obtained of the true probability of rejecting the null hypothesis at these significance levels in this null situation. Each run was done with $n = 100$ for cases and controls, and for each specific situation 1000 sets of data were generated.

The main conclusion to be drawn from these results is that overestimation of significance can be very pronounced, that it increases with increasing correlation of X_3 and (X_1, X_2), and with more overlap of cases and controls. For values of ρ less than 0·5 (multiple correlation coefficient of X_3 and (X_1, X_2) is $\sqrt{2\rho} = 0{\cdot}71$) the error is not large.

In case-control studies in which many variables are incorporated in the regression analysis (that is, large k), the multiple correlation of X_k and $(X_1 \ldots X_{k-1})$ will tend to be large. Values of k up to 20 or more have been used. This is precisely the situation in which the significance level claimed for X_k using the multivariate confounder score is most likely to be much exaggerated.

Comparison with multiple logistic regression

Miettinen's approach has also been suggested for the situation in which logistic regression rather than standard regression is more applicable. This leads, however, to the same exaggerated significance levels.

The regression function takes the form

$$\Pr(\text{Case} \mid X_1 \ldots X_k) = 1 - \Pr(\text{Control} \mid X_1 \ldots X_k) = \exp(b_o + b_1 X_1 + \ldots + b_k X_k) \,/\, [1 + \exp(b_o + b_1 X_1 + \ldots + b_k X_k)] \qquad (11)$$

Write the likelihood function generated by the observed x's as $L(b_o \ldots b_k)$. The maximum likelihood values $(\hat{c}_o \ldots \hat{c}_k)$ maximise L and are equivalent to the $\hat{c}$'s in (4). For the equivalent coefficients to (5) we maximize L subject to $b_k = 0$—write the resultant coefficients as $(\hat{d}_o \ldots \hat{d}_{k-1}, 0)$. The standard likelihood ratio test for significance of the $\hat{c}_k$ is given by

$$2[\ln L(\hat{c}_o \ldots \hat{c}_k) - \ln L(\hat{d}_o \ldots \hat{d}_{k-1}, 0)] \qquad (12)$$

which, on the null hypothesis of $b_k = 0$, follows approximately a χ^2 distribution on 1 degree of freedom.

In place of (12), however, the confounder score method test is approximated by

$$2[\ln L(\hat{c}_o \ldots \hat{c}_k) - \ln L(\hat{c}_o \ldots \hat{c}_{k-1}, 0)] \qquad (13)$$

The estimate of b_k which results from the use of Miettinen's confounder score is again correct, but expression (13) will always be larger than (12), and the significance test (13) will therefore overstate the significance of the value $\hat{c}_k$. Quantitatively, this overstatement will be similar to that seen in the Table, because normal discriminant and logistic discriminant results are usually close.

Table *Simulation estimates of probability (%) of exceeding upper 1-sided nominal significance level**

		1-sided nominal significance level (%)			
$\mu_1 = \mu_2$	ρ	*5*	*2·5*	*1*	*0·5*
0·5	0·7	27·0	22·3	17·9	15·3
0 5	0·65	13·1	9·8	8·2	6·7
0·5	0·6	8·4	5·8	3·4	2·3
0·5	0·5	6·5	3·1	1·7	0·9
0·5	0·3	4·7	1·9	0·9	0·8
0·5	0·0	5·8	3·1	1·2	0·5
1·0	0·7	21·5	19·1	16·4	14·7
1·0	0·65	7·8	5·0	2·8	2·1
1·0	0·6	6·8	3·6	1·6	1·0
2·0	0·7	15·4	13·3	11·8	10·6
2·0	0·65	7·8	4·2	2·5	1·6
2·0	0·6	5·4	3·6	1·4	0·7

*n = number of cases = number of controls = 100.
Each line of results is based on 1000 simulations.

Reprints from Professor M. C. Pike, Edmondson Research Building, 1840 N. Soto Street, Los Angeles, CA 90032.

References

Anderson, T. W. (1957). *An Introduction to Multivariate Statistical Analysis.* Wiley: New York.

Breslow, N., and Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics,* **34,** 100–105.

Holford, T. R., White, C., and Kelsey, J. L. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology,* **107,** 245–256.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute,* **22,** 719–748.

Miettinen, O. S. (1976). Stratification by a multivariate confounder score. *American Journal of Epidemiology,* **104,** 609–620.